SOME NOVEL MEASUREMENT ON EXPLORING LARGE DATA SETS BASED ON MULTI-VARIABLES MUTUAL INFORMATION THEORY

[1,2]YU-SHAN JIANG, [1] QING-LING ZHANG AND [2]CHAO LIU

[1] State Key Lab. of Integrated Auto. of Process Ind., Northeastern Univ.Shenyang, Liaoning, China

[2] School of Mathematics and Statistics, Northeastern Univ. at QinhuangdaoQinhuangdao, Hebei, China

E-mail: [1]sobolev@126.com, [2]jys@mai.neuq.edu.cn

ABSTRACT

Applying for information theory, we present a measure of dependence for three-variable relationships: the three variables maximal information coefficient (3D-MIC). It is a kind of maximal information-based nonparametric exploration (MINE) statistics for identifying and classifying relationships in large data sets. 3D-MIC generalized the MIC measurement. At the same time some optimal single axis partition algorithm (OSPA) is built to ensure the feasibility of the MIC measurement.

Keywords: *Mutual Information, Maximal Information-Based Nonparametric Exploration (NE), Dynamic Programming, Maximum Information Coefficient (MIC), Optimal Single Axis Partition Algorithm (OSPA)*

1. INTRODUCTION

Data sets of large size are increasingly common in fields as various as genomics, physics, political science, and economics. Exploring large data sets to discover relationships among variables becomes important and is full of growing challenges. If you do not already know what kinds of relationships to search for, how do you efficiently identify the important ones? One way to begin exploring a large data set is to search for several pairs of variables that are closely associated. To do this, we could calculate some measures of dependence for each pair, rank the pairs by their scores, and examine the top-scoring pairs. It belongs to a larger class of maximal information-based nonparametric exploration (MINE) statistics for identifying and classifying relationships.

The nonparametric methods, such as the tests, the Fisher exact probability test, and the Spearman rank correlation, have long been among the standard tools of the statisticians [2]. Recently, some new nonparametric or 'distribution-free' statistical methods [3-10] have gained prominence in statisticians. In [3] a detailed analysis of the LPA algorithm based on local information is done which improved the LPA-the SNA algorithm. In skew correlation Deng [4] developed a method called Based Skew Double Triangle algorithm to study the correlation between AT and CG. García [5] discuss some multiple comparisons nonparametric approaches with the computation of adjusted p-value which improve the results offered by the Friedman test in some circumstances. By using Maximum Likelihood Tamura [8] analyze the molecular sequence data which improves the computational efficiency and the accuracy of the estimates. Jishnu [9] performs a systematic analysis of interaction dynamics across different technologies and shows the high-throughput yeast two-hybrid is the only available technology for detecting transient interactions on a large scale.

Our works are derived from [1]. David N. Reshef and his work team detect a novel association in large data sets which is called the maximal information coefficient (MIC) measure of dependence for two-variable relationships. They have proved that the MIC of a noiseless functional relationship converges to 1 as sample size grows. And if the sample distribution (X, Y) is statistically independent, the MIC converges to zero. Intuitively, MIC is based on the idea that if a relationship exists between two variables, then a grid can be drawn on the scatter plot of the two variables that partitions the data to encapsulate that relationship.

The organization of the study is as follows. In section 2 we present some three variables mutual information which generalized the common mutual information and establish the definition of 3D-MIC in three statistical distributions (X, Y, Z). In section 3 we present the single axis partition algorithm to approach the 3D-MIC. Based on heuristic dynamic programming algorithm the recurrence algorithm is efficiency for given parameter $B(n)$ in practice.

2. PRELIMINARIES AND DEFINITIONS

There are some definitions and lemmas that we will use in the following section. Firstly, we show some concepts of information given in this theory which include entropy, joint entropy, conditional entropy, relative entropy, mutual information.

2.1 Some Information Theory Conceptions

Suppose X, Y, Z are three statistic variables in X, Y, Z respectively. Let $p(x)=\Pr(X=x)$ be the probability of X=x. The entropy of X is defined by

$$H(X)=-E[\log p(x)]=-\sum_{x\in X}p(x)\log p(x)$$

The joint entropy of X, Y is defined by

$$H(X,Y)=-E[\log p(x,y)]=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(x,y)$$

The conditional entropy is defined by

$$H(Y\mid X)=-E_{p(X,Y)}[\log p(Y\mid X)]=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(y\mid x)$$

The relative entropy is

$$D(p\,\|\,q)=E_{p(X,Y)}[\log\frac{p(X)}{q(X)}]=\sum_{x\in X}p(x)\log\frac{p(x)}{q(x)}$$

The mutual information is defined by

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

From above one can obtain

$$I(X;Y)=H(X)+H(Y)-H(X,Y)$$

Now, let us denote the mutual information in three variables

$$\begin{aligned}I(X;Y;Z)&=H(X)+H(Y)+H(Z)\\&\quad-H(X,Y)-H(Y,Z)-H(X,Z)\\&\quad+H(X;Y;Z)\end{aligned}$$

(1)

The symmetry of the joint entropy H(X, Y) = H(Y,X) implies that the mutual information in three variables remains symmetry.

2.2 Main Definition on 3D-MIC

Given a finite set $D=\{(x,y,z)\mid x\in X, y\in Y, z\in Z\}$ of three variables, we can partition the x-value of D into x bins, the y-value of D into y bins and the z-value of D into z bins allowing empty bins. We call such a partition an x-by-y-by-z grid G. Given a grid G, let $D|_G$ be the distribution induced by the points in D on the boxes of G, that is, the distribution on the boxes of G obtained by letting the probability mass in each box be the fraction of points in D falling in that box.

For a fixed D, different grids G result in different distributions $D|_G$. To exploit this fact in defining 3D-MIC, we first of all make the following definitions.

Definition 1 For a finite set $D\subset\mathbb{R}^3$ and positive integers *x*, *y*, *z*, define

$$I^*(D,x,y,z)=\max I(D|_G) \qquad (2)$$

where the maximum is over all grids *G* with *x*-by-*y*-by-z grid, and $I(D|_G)$ denotes the three variables mutual information of $D|_G$.

Definition 2 The characteristic matrix $M(D)$ of a set *D* of three-variable data is an infinite matrix with entries

$$M(D)_{x,y,z}=\frac{|I^*(D,x,y,z)|}{\log(xyz)} \qquad (3)$$

Definition 3 The *Three Dimensions Maximal Information Coefficient* (*3D-MIC*) of a set *D* of three-variable data with sample size *n* and grid size less than $B(n)$ is given by

$$3D-MIC(D)=\max_{xyz<B(n)}M(D)_{x,y,.z} \qquad (4)$$

where $B(n)\le O(n^{1-\varepsilon})$ for some ε .

Remark Since for an *x*-by-*y*-by-z grid *G*, $0<|I(D|_G)|<\log(xyz)$, all the characteristic matrices fall between 0 and 1. The symmetry of 3D mutual information implies that the characteristic matrix $M(D)$ remains the same when the axis of *D* are interchanged. However, computing each entry of the characteristic matrix grows exponentially with the number of data points increasing. The maximum 3D mutual information on Data sets is only one kind of statistic property. There can be other different patterns in definition of corresponding statistic scores and measurement.

3. RECURRENCE APPROACH ON 3D-MIC ALGORITHM

We begin by outlining an idealized algorithm for generating the characteristic matrix. Algorithm 1 represents what we would use if efficiency were not a problem.

Table 1: Algorithm Of Compute Characteristic Matrix

```
Algorithm 1 Characteristic Matrix.(D, B)
Require: D is a three variable set
for ( x, y, z) such that xyz < B do
    $G \leftarrow (x, y, z)$ grid on D
    $I^*(D, x, y, z) \leftarrow \max \mathrm{I}(D|_G)$
    $M(D)_{x,y,z} = |I^*(D, x, y, z)| / \log(xyz)$
end for
return $M(D)_{x,y,z} : xyz \le B$
```

By definition (1), the maxi function involved in Algorithm 1 is meant to return the highest mutual information attainable using a grid G with x−by−y−by−z on the data D. The core of approximating maxi is to find an optimal single axis partition. In the following section, we will use some dynamic programming methods which are called optimal single axis partition (OSP) to obtain the maxI.

Assume that our set $\{(x, y, z)|(x, y, z) \in D\}$ is sorted in an increasing order by x-value, we denote various partitions of the x-axis by specifying the indices of the end points of their columns. Specifically, we will call an ordered list of integers $\langle p_0, \cdots, p_t \rangle$ with $p_0 < p_1, < \cdots, < p_t$ an x-axis partition of size t of the $(p_0 + 1)$-th through p_t-th points of D. Given a partition $P = \langle p_0, \cdots, p_t \rangle$ and an integer a with p_i < a < p_{i+1} , we denote $P \bigcup \{a\} =: \langle p_0, \cdots, p_i, a, p_{i+1}, \cdots, p_t \rangle$. If p_i =a for some i, we denote $P \bigcup \{a\} = P$, and if a > p_t , we denote $P \bigcup \{a\} :=< p_0, \cdots, p_t, a >$. For fixed y, z-axis partitions Q, R, by definition (1) on mutual information in three variables, we have

$$\begin{aligned} I(P;Q;R) &= H(P) + H(Q) + H(R) \\ &\quad - H(P,Q) - H(P,R) - H(Q,R) \\ &\quad + H(P,Q,R) \end{aligned}$$

where $H(\cdot)$ denotes Shannon entropy. However, since Q, R are fixed, the OSP algorithm needs only to maximize

$$I(P;Q;R) = H(P) - H(P,Q) - H(P,R) + H(P,Q,R)$$

over all partitions P on x-axis of D in order to maximize *I*(*P*,*Q*,*R*). Thus, from this point of view we show the following optimal single axis partition algorithm to approach 3D-MIC Algorithm.

Table 2: Single Axis Partition Algorithm

```
Algorithm 2 3D-MIC Algorithm (D, Q, R, x)
Require: D is a three variable ( x, y, z,) set sorted in
 increasing order by x-value
Require: Q is a y-axis partition of D
Require: R is a z-axis partition of D
Require: x is a integer greater than 1
Ensure: Returns a list of scores $(I_2, \cdots, I_x)$ such that each
 $I_l$ is the maximum value of I (P; Q; R) over all partitions P on x-axis of size l.
$< p_0, \cdots, p_k > \leftarrow$ Get Partition (P; Q; R)
%% Find the optimal partitions of size 2
for t=2 to k do
    Find $s \in \{1 , \dots , t\}$ maximizing
    $H(< p_0, p_s, p_t >) - H(< p_0, p_s, p_t >, Q)$
    $- H(< p_0, p_s, p_t >, R) + H(< p_0, p_s, p_t >, Q, R)$
    $P_{t,2} \leftarrow < p_0, p_s, p_t >$
    $I_{t,2} \leftarrow H(P_{t,2}) + H(Q) + H(R)$
    $- H(P_{t,2}, Q) - H(P_{t,2}, R) - H(Q, R) + H(P_{t,2}, Q, R)$
end for
%% Inductively build the rest of the table of
  %% optimal partitions
for l=3 to x do
    for t=l to k do
        Find $s \in \{l\text{-}1 , \dots , t\}$ maximizing
        $F(s,t,l) := \frac{p_s}{p_t}[I_{s,l-1} - H(Q) - H(R) + H(Q,R)]$
        $-\frac{p_t - p_s}{p_t}[H(< p_s, p_s >, Q) + H(< p_s, p_s >, R)$
        $- H(< p_s, p_s >, Q, R)]$
        $P_{t,l} \leftarrow P_{s,l-1} \bigcup p_t$
        $I_{t,l} \leftarrow H(P_{t,l}) + H(Q) + H(R)$
        $- H(P_{t,l}, Q) - H(P_{t,l}, R) - H(Q, R) + H(P_{t,l}, Q, R)$
    end for
end for
$P_{k,l} \leftarrow P_{k,k}$ for $l \in (k , x]$
$I_{k,l} \leftarrow I_{k,k}$ for $l \in (k , x]$
return $I_{k,2}, \cdots, I_{k,x}$
```

If given some number x of partitions and some number y, z of partitions, we could run the OSP algorithm function on every possible y, z size partition. we would find an optimal grid. But the number of possible y, z partitions makes this infeasible. A natural approach to this problem is to consider only grids for which at least one axis is partitioned ahead. To this end, the OSP algorithm fixes a partition of y, z axis with Q, R and then runs to the result. Later, the OSP is called again but with the axes switched. The maximum of the three scores obtained is used.

4. CONCLUSIONS

This study describes the improved 3D-MIC Algorithm that discovers coefficients in multi-variables (X, Y, Z). We have showed that the 3D-MIC Algorithm is different from the original one in complexity and properties. The idea in 3D-MIC can be generalized to n-dimension variable sets. Like any other statistical analysis, the proposed approach has some limitations. First, it cannot automatically determine the maximum grid numbers B(n). B(n) is the important factor affecting program running speed. Furthermore, we need to give a complete proof of the convergence of the algorithm theoretically. This is our follow-up work.

5. ACKNOWLEDGMENTS

The research is supported by N.N.S.F. of China under Grant No. 61273008 and No. 61104003. The research is also supported by the Key Laboratory of Integrated Automation of Process Industry (Northeastern University).The author is grateful to the anonymous referee for a careful checking of the details and for helpful comments that allow us to improve the manuscript.

REFERENCES

[1] David N. Reshef, "Detecting Novel Associations in Large Data Sets", Science, Vol. 334, 2011, pp. 1518-1524.

[2] Sidney Siegel, “Nonparametric Statistic”, The American Statistician, Vol. 11, No. 3, 1986, pp. 13-19.

[3] Sheng Xin, "The research on LPA algorithm and its improvement based on partial information", Journal of Theoretical and Applied Information Technology, Vol. 43, No. 2, pp. 192-197.

[4] Xuegong Deng, Ilkka Havukkala and Xuemei Deng, "Large-scale genomic 2D visualization reveals extensive CG-AT skew correlation in bird genomes", BMC Evolutionary Biology, Vol. 7, No.234, 2007, pp. 1471-1482.

[5] Stéphane Guindon, Jean-François Dufayard, Vincent Lefort, etc, "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0", Systematic Biology, Vol. 59, No. 3, 2010, pp. 307-321.

[6] Salvador García, Alberto Fernández and Julián Luengo, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power", Information Sciences, Vol. 180, No. 10, 2010, pp. 2044-2064.

[7] Piotr Kulczycki, "Nonparametric Estimation for Control Engineering", 4th WSEAS/IASME International Conference on Dynamical Systems and Control, 2008,Vol. 1, pp. 115-121.

[8] David M. Erceg-Hurn, Vikki M. Mirosevich, "Modern Robust Statistical Methods", American Psychological Association, Vol. 63, No. 7, 2008, pp. 591-601.

[9] Koichiro Tamura, Daniel Peterson, Nicholas Peterson, "MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance and Maximum Parsimony Methods", Molecular Biology and Evolution, Vol. 28, No. 10, 2011, pp. 2731-2739.

[10] Jishnu Das, Jaaved Mohammed and Haiyuan Yu, "Genome-scale analysis of interaction dynamics reveals organization of biological networks", Bioinformatics, Vol. 28, No. 14, 2012, pp. 1873-1878.

[12] T.M. and Thomas, J. A., "Elements of Information Theory", John Wiley \& Sons, Inc., New York, USA. 1991.